A review on Neural Signal Compression methods

Samira Riki*
*Graduated Master Student, Electrical and Computer Engineering Department, K.N. Toosi University of Technology, Tehran, Iran

ABSTRACT
There are many techniques for neural data reduction that follow different factors. Neuroscientists are so interested in low power and also simple types, especially in intra-cortically applications that the electrodes are directly implanted into the cortical and also in Brain-Machine Interface (BMI) systems. In this work at first I describe about the intra cortical systems and I introduce their mechanism. At continue some of neural data reduction methods are introduced that are suitable for the intra-cortically applications. The neuroscientists worked on this methods and tried to improve the key factors such as power consumption and bit rate. In this way, they accessed to some achievements in different factors. But the notable point is that every methods reached to different achievements and also none of techniques can improve the whole factors, so they tried to set a trade-off between the key factors. So it can be said that all of this techniques are suitable for different applications and they should be chosen by considering the application.

Key words:
data reduction, neural signal, implantable microsystems, intra-cortically microsystems, BMI systems

1. Introduction

Multi-Electrode Arrays (MEAs) are systems for activity recording and stimulating up to hundreds of neurons which are released in implantable MEAs and non-implantable MEAs. Implantable types are beneficial for intra-cortical applications. The recorded information are useful in systems called Brain-Machine Interfaces (BMIs). BMIs are systems that provide communication between the brain and an external technology by converting the neural signals to actions [1]. Therefore these systems are appropriate for patients who suffer diseases such as epilepsy, Parkinson and the other neurological diseases [2][3]. These systems are available in wireless and wired types. The wireless types by reduction the post-surgical risks and portability, are a good alternative to wired types [4]. In design and development of these systems there are considerable constraints such as small size of silicon area and limitation of wireless bandwidth and power consumption. Neuroscience researchers are enthusiastic to increase the number of simultaneous recording channels as much as possible to promote the accuracy and functionality of the BMIs/BCIs systems furthermore because of some design challenges they are interested in data reduction in such a way that the most significant information being sent instead of sending the whole information.

Recording the electrical activities of nervous cells is called electrophysiological and it is consists of two kinds: non-invasive approaches such as EEG, MEG and invasive approaches such as ECoG, Intra-cortical. The invasive kinds are more complex than the others because in this approaches electrodes are implanted on a part of cortex or implanted intra cortex. This techniques have a good accuracy and they have higher time-resolution and spatial-resolution in comparison to the other techniques. Also the intra-cortical electrodes are portable. Portable means that the electrodes are carried by the patient. So in this work I concentrated on intra-cortically approaches.

1.1 Intra-cortical neural recording system

Figure 1 illustrates an intra-cortical neural system. This system is consists of an implantable part and an external setup. Implantable cortical neural recording, in general, consist of two building block [5]:

- **Neural Recording front-end module**
  In this module neural signal recording, preparing and multiplexing is done.
- **Main module**
  Providing wireless transmission (such as data and power consumption), encoding, and decoding, and data reduction in this module is took placed.

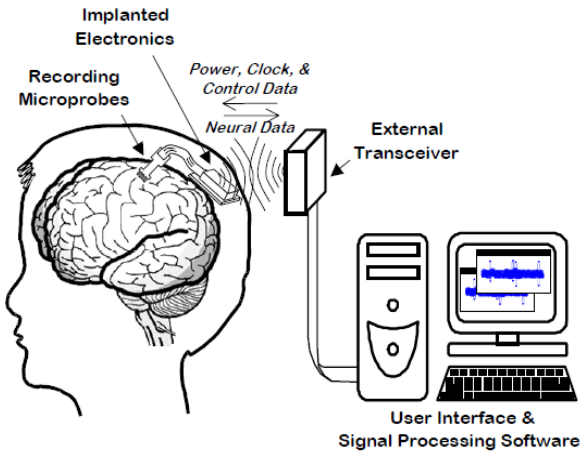

Figure 1. The neural recording microsystem [6]

Manuscript received February 5, 2017
Manuscript revised February 20, 2017

2. Challenges In The Increasing The Density Of Recorded Channels

BMI systems applies many constraints to wireless transmission since more than 100 electrodes are sending neural data and in future this number increases about 1000 electrodes.

In a wireless neural recording microsystem, as the number of recording channels increases, many issues are showed off. One of the most important design challenges is power consumption. It is very necessary to decrease the power consumption due to decrease the number of charge and discharge times. Studies have shown that if the chip leads the brain tissue to temperature more than about 1°C the brain tissue will be damaged, also power consumption density of more than 2mW / cm 80 for tissue would be harmful.

Another constraint is wireless bandwidth while in a wireless neural recording system if 100 recording channel transmit data in 25kHz sampling rate and 8-bits of resolution then generated data is about 20Mb/s. Transmitting this value of information by a wireless link is infeasible and also researches have shown that the silicon area should be less than 1cm2 [7].

3. Neural Data Reduction Approaches

There are different ways to compress data and they can be categorized in different classes. This approaches are one of two types: Lossless compression and lossy compression. Through compression, the order of input samples converts to a new order and the new arrangement is shorter than main order. In lossless approaches reconstructed signal is equivalent to original signal and information is saved completely whereas in lossy approach, reconstructed signal is an approximation of original one and the difference between the two is called the error signal compression. In this approach unimportant parts of signal are discarded, so more compression rate is expected. This method is used in compression of images and audio encoding. If the key characteristics of the neural signal being extracted, Lossy methods can be an appropriate choice for neural data compression. The methods of neural signal compression are generally lossy types. Figure 2 illustrates a division of lossy compression approaches. Recently, there are three categories of lossy compression: Event-based approaches, transformation-based approaches, and Compressed Sensing approaches [8]. An example of event based approaches is spike detection. In these methods one or two bits-pulse are reported as a spike event so in these methods the spike shapes are not transmitted. If the whole signal is required, the transformation-based approaches are usually selected. Recently, the field of Compressed Sensing (CS) has shown capability in neural signal compression. This method has a simpler circuit in comparison to transformation-based methods.

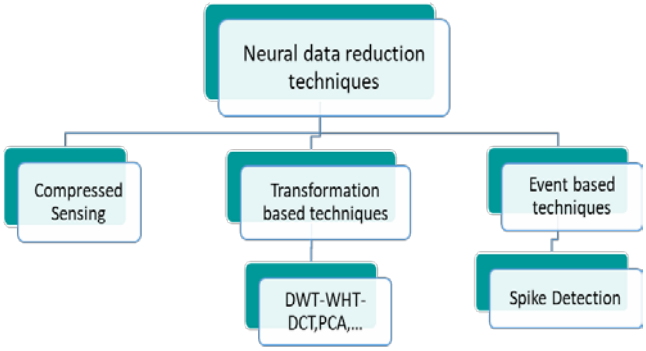

Figure 2. Flowchart of the classification of neural data reduction techniques

3.1. Spike detection

A neural signal is consist of an action potential about 100 to 500 µV and background noise and some low-frequency parts. So at the first step neural signal should be amplified before spike detection to be prepared for this stage. So the signal is amplified with a gain around 40 to 60 dB and also the signal is filtered and its low-frequency (below 1~ 10Hz) and high frequency (above 7~ 10 kHz) parts are filtered out. Then this signal is delivered to the spike detection stage.

In general, an intra-cortically neural signal consists of three main components: action potentials (known as spikes), local field potential (LFP), and background noise. The action potential can be the part of signal that has more amplitude in comparison to the background noise because of the voltage difference among signal and noise. There are various spike detection approaches that can be categorized into two classes: Feature-based spike detection approaches and spike detection by hard thresholding. In the first approach, a preconditioner searches for some certain features to detect the spike occurrence. In the second approach a threshold value is considered, and while the signal value be more than this value the spike occurrence is reported. In some of these approaches, only the features of action potential are transmitted to the external setup. So for each channel one or two bits are transmitted as channel status [9] or active channel addresses [6]. On the other hand, in this approach one or multi sparse bits are transmitted instead of sending the whole signal, so the transmit bandwidth considerably decreased. Also in some of neural reduction techniques, spike detection is the first step. In [10] a processor has been introduced that assumes a threshold voltage, then the value of signal is compared to the threshold value so the

samples more than threshold value are packaged as their address and then they are sent. This system is real-time but because of sending 5-bits address for 5-bits sample it sends many information. The mentioned spike detection ASIC occupies 6mm2 in 0.5µm technology and consumes 2.6mW of power at 3-V supply. Although the spike detection methods contain useful information, but in these methods the signal waveform is discarded, so these methods are not applicable for applications that require the whole signal waveform. In order to maintain signal waveform in [11-13], samples are compared with the threshold value samples are buffered then when spike is detected, a burst of action potential samples are sent.

Intra-cortically detection methods have benefits such as high accuracy and simplicity but one of the most challenges of these methods is action potentials independency. There are different approaches for spike detection such as threshold level, and wavelet transform, and energy evaluation. This approaches can be analog or digital. In application that have low resolution (to 9-bits) the analog approaches are preferred.

3.2. Transformation based data reduction

Mathematical transforms are among the most common methods for data compression. In addition to compression, these methods preserve wave shape of the action potential. At continue some of more functional approaches such as Discrete Wavelet Transform (DWT), and Discrete Cosine Transform (DCT), and Walsh-Hadamard Transform (WHT) are described.

3.2.1. DWT

Recently the Discrete Wavelet Transform (DWT) has a successfully leap on compression of neural information [14-16]. The DWT transforms discrete signal and convert it from the time domain into the frequency domain. To achieving a level of DWT at first some high-pass filters and low pass filters are applied. Then data sampling is applied on filtered data to obtain detail and approximation. In general, for neural signal compression, it is most optimal if select such a wavelet function to estimates signal by minimum coefficients and error. It has been shown that the most energy of the signal is concentrated in a few large coefficients while many small coefficients carry insignificant information such as noise. So, to achieve high data reduction rates, the coefficients are filtered through a thresholding stage. Also the threshold level should be set carefully. Because of constraints such as power and size of silicon area in implantable devices, VLSI implementation of these systems has a great importance. In [14] it has be shown that the function of symmlet4 because of the similarity to the action potential wave shape, is appropriate for neural signal compression. A 32-channel compression system is proposed in [15] based on 4-level symmlet4 DWT. This system compresses data more than 20 times in sampling rate of 25 kHz and by resolution of 10 bits/sample. Also it occupies 5.75 mm2 in 0.5µm CMOS technology and consumes 3mW of power. In [16] a compression microsystem is proposed that it works based on the Discrete Haar Wavelet Transforms (DWHT). From the standpoint of compression, the haar basis functions cannot have a good performance like symmlet and Daubechies. But from standpoint of hardware implementation, it is simple and by considering the constraints of implementation, increasing of recording channels can be done. In the mentioned work, the hardware of compression system for only two points DHWT consists of a buffer, an adder, and a subtraction. In the mentioned work, for comparison Haar to Symmlet4 basis functions both approaches are designed for a single channel. So results illustrate that the error value for DWHT is just 0.01% more than symmlet4 case. On the other hand, by haar wavelet 83% decrement of transistor and about 90% decrement of size in 130 nm of technology is improved. The 64 channel of this approach compress neural data with compression ratio of 112 and error of near 2%. This system occupies 0.1mm2 and it consumes about 0.12mW at 1.2V supply voltage. Therefore, it can be said that in implantable applications the improvement in hardware simplicity and power consumption is much more significant than the reconstruction error. However, it can be said that the appropriate approach should be selected by considering the application

3.2.2. DCT

Discrete cosine transform (DCT) represents finite sequence of points as the sum of cosine functions with different frequencies. In this method, instead of the signal, cosine series coefficients represent the information. On the other hand, data is converted from time space to frequency space. Also this transform is usually used in photo compression. The DCT transform obtains by:

A 128-channel compression module of the proposed approach in [17] compresses neural signal with a compression ratio of 69 and Root-Mean Square (RMS) of 6 %.

$$Y(\mathrm{k}) = \sqrt{\frac{2}{N}} a_k \sum_{n=0}^{N-1} x(\mathrm{n}) \cos\left(\frac{(2\mathrm{n}+1)\mathrm{k}\pi}{2N}\right), k = 0,1,...,\mathrm{N}-1 \qquad (1)$$

Also the reverse of this transform is:

$$X(\mathrm{n}) = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} a_k \, \mathrm{Y}(\mathrm{k}) \cos\left(\frac{(2\mathrm{n}+1)\mathrm{k}\pi}{2N}\right), k = 0,1,...,\mathrm{N}-1 \qquad (2)$$

In the equations x(n) is input signal and N is the number of samples in time domain, Y(k) is transformed signal, k is the number of coefficients in DCT domain, a_k is a constant

value that is equal to one for k=0 and otherwise it is equal to 1/√2. The DCT is an orthogonal transform and its basis functions are sinusoidal. The transform is used in applications such as compression of pictures and audio. It can be because of concentration the considerable information of signal to a few large coefficients. In [17] an optimal approach with low multiplier for neural signal compression is introduced. The approach is consists of several stages. At the first stage, signal is taken to the DCT space and coefficients are produced. In continue, the coefficients are compared with a threshold value and the whole of under threshold values are replaced by zero. Then Run Length Encoding (RLE) is applied on the coefficients. After this stage, all of non-zero coefficients are directly sent to the external setup and instead of sending all of zero values, just the number of zero values is reported, therefore the submitted data is decreased. In the receiver side, Run Length Decoding (RLD) and the reverse of DCT is applied on the received coefficients and the signal is reconstructed.

A 128-channel compression module of the proposed approach in [17] compresses neural signal with a compression ratio of 69 and Root-Mean Square (RMS) of 6 %.

3.2.3. WHT

The Walsh-Hadamard Transform (WHT) is one of the mathematical transforms. The transform vector is a vector with 2n ×2n dimensions that H0=1 and the other elements are 1 or -1. The recursive Walsh-Hadamard matrix is:

$$H_m = \frac{1}{\sqrt{2}}\begin{bmatrix} H_{m-1} & H_{m-1} \\ H_{m-1} & -H_{m-1} \end{bmatrix}, H_0 = 1$$

H_2 is a two order Fourier transform. The second order (H_2) is used in [18]. One advantage of this transform is that it doesn't need to multiplier and divider. In [18] this transform is used for neural signal compression. Figure 3 illustrates the distribution of energy in the neural signal after applying the WHT, so it can be said that the considerable segment of energy is aggregated on a few large coefficients.

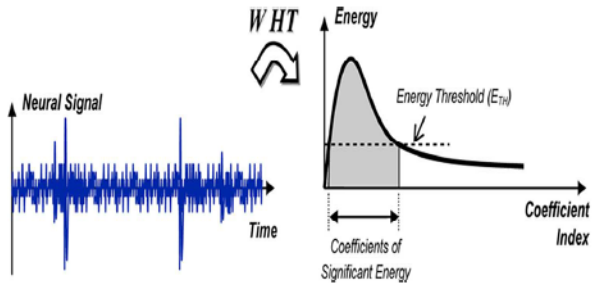

Figure 3. Illustration of energy concentration of neural signals after applying WHT [18].

In [18] a coprocessor is introduced that it compresses the neural signal with compression ratio of 63 times and RMSE of 4.66 %. The 128-channel of proposed approach in 180 nm technology occupies about 1.64 mm^2 and it consumes near 81 µW in 250 kHz of sampling kHz and at supplying voltage of 1.8 V.

3.3. Compressed Sensing

One of the approaches that have been proposed in recent years for data compression is compressed sensing approach that attracted the attention of many researchers and scientists in the field of neuroscience. Compressed Sensing approach (CS) applies to a signal processing technique for signal compression. This technique, like other compression methods tries to send more information by paying less penalty. Based on the Nyquist theorem in order to reconstruct a sampled signal, sampling frequency should be at least twice the maximum frequency. On the other hand, in many applications, because of bandwidth limitations in wireless transmission some parts of the data is discarded. So this idea comes to mind that instead of sampling in Nyquist rate and then discarding some data by doing a transform such as DWT, DCT, the sampling and compression being applied simultaneously. This idea is the same Compressed Sensing (CS). In CS instead of sampling the signal, the signal is measured. Measure is a linear combination of multiple samples. Then, using the measurements, the reconstructed signal is made. It is obvious that the number of measurements required for signal reconstruction in CS is less than the number of samples required for signal recovery based on Nyquist's theorem. CS is used for sparse signals. The Sparse signal is a signal that the most samples of it is zero or near to zero. So this approach is appropriate for audio compression while neural signals are not necessarily sparse in time domain. So this approach cannot be used directly in neural signals therefore a sparse applying stage is needed. So far, a lot of research in the field of compressed sensing is provided, also it is tried to choose a sensing matrix in order to have a simple circuit and to be optimal in signal recovery. In [19] a Minimum Euclidean Matrix or Manhattan Distance Cluster-based (MDC) is chosen as a sensing matrix. In the approach the neural signal is compressed with reconstruction error being around 0.2 and it achieves a compression rate up to 90%. The proposed system consumes 0.59mW and it occupies 7µm^2. Results show as a sensing matrix be more similar to neural signal, may achieve more compression ratio. Based on this feature, in [20] an approach is proposed that it concentrates on improvement of compression ratio and hardware. This system occupies an area around 200µm × 300µm per recording channel in 180 nm and consumes 0.27µW at 20 kHz.

Conclusion

There are different approaches for signal compression but based on mentioned constraints, all of the approaches may not be appropriate for neural signals. To overcome the limitation of bandwidth a wide variety of compression techniques are reported. These techniques are one of the spike reporting, mathematical or CS techniques. From standpoint of bandwidth saving, the spike detection approaches are successful but this approaches discard the wave shape. So the mathematical approaches can transmit the wave shape, also these techniques from the standpoint of data compression have been successful. On the other hand, increasing the number of recording channels turned to a challenge for neuroscientists. Furthermore they try to increase the number of channels in such a way that the bandwidth limitation and power consumption be considered. In addition to the described approaches in this work, in [5] some hardware techniques are described that result in considerable reduction of bit-rate in multi-channel neural recording microsystems. In overall it cannot be said that what approach is better than the other, but it should be considered that for what application the approach is used and how much power consumption and bit-rate is allowed.

References

[1] Muhammad Naeem, T. (2015). Architecture and system level concept for wireless brain machine interface.

[2] Li, N., & Sawan, M. (2015, April). High compression rate and efficient spikes detection system using compressed sensing technique for neural signal processing. In Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on (pp. 597-600). IEEE.

[3] Lapolli, A. C., Coppa, B., & Heliot, R. (2013, July). Low-power hardware for neural spike compression in BMIs. In Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE (pp. 2156-2159). IEEE.

[4] Thorbergsson, P. T., Garwicz, M., Schouenborg, J., & Johansson, A. J. (2014). Strategies for high-performance resource-efficient compression of neural spike recordings. PloS one, 9(4), e93779.

[5] Judy, M., Akhavian, A., & Asgarian, F. (2015). Data Reduction Techniques in Neural Recording Microsystems.

[6] Sodagar, A. M., Wise, K. D., & Najafi, K. (2007). A fully integrated mixed-signal neural processor for implantable multichannel cortical recording. IEEE Transactions on Biomedical Engineering, 54(6), 1075-1088.

[7] Rogers, C. L. (2007). Ultra-Low power analog circuits for spike feature extraction and detection from extracellular neural recordings (Doctoral dissertation, University of Florida).

[8] Suo, Y., Zhang, J., Xiong, T., Chin, P. S., Etienne-Cummings, R., & Tran, T. D. (2014). Energy-efficient multi-mode compressed sensing system for implantable neural recordings. IEEE transactions on biomedical circuits and systems, 8(5), 0-0.

[9] Harrison, R. R., Watkins, P. T., Kier, R. J., Lovejoy, R. O., Black, D. J., Greger, B., & Solzbacher, F. (2007). A low-power integrated circuit for a wireless 100-electrode neural recording system. IEEE Journal of Solid-State Circuits, 42(1), 123-133.

[10] Olsson, R. H., & Wise, K. D. (2005). A three-dimensional neural recording microsystem with implantable data compression circuitry. IEEE Journal of Solid-State Circuits, 40(12), 2796-2804.

[11] Gosselin, B., Ayoub, A. E., Roy, J. F., Sawan, M., Lepore, F., Chaudhuri, A., & Guitton, D. (2009). A mixed-signal multichip neural recording interface with bandwidth reduction. IEEE Transactions on Biomedical Circuits and Systems, 3(3), 129-141.

[12] Gosselin, B., & Sawan, M. (2010). A low-power integrated neural interface with digital spike detection and extraction. Analog Integrated Circuits and Signal Processing, 64(1), 3-11.

[13] Bonfanti, A., Ceravolo, M., Zambra, G., Gusmeroli, R., Spinelli, A. S., Lacaita, A. L., ... & Fadiga, L. (2010, August). A multi-channel low-power system-on-chip for single-unit recording and narrowband wireless transmission of neural signal. In Engineering in medicine and biology society (EMBC), 2010 annual international conference of the IEEE (pp. 1555-1560). IEEE.

[14] Oweiss, K. G. (2006). A systems approach for data compression and latency reduction in cortically controlled brain machine interfaces. IEEE Transactions on Biomedical Engineering, 53(7), 1364-1377.

[15] Kamboh, A. M., Oweiss, K. G., & Mason, A. J. (2009, May). Resource constrained VLSI architecture for implantable neural data compression systems. In Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on (pp. 1481-1484). IEEE.

[16] Shaeri, M. A., Sodagar, A. M., & Abrishami-Moghaddam, H. (2011, August). A 64-channel neural signal processor/compressor based on Haar wavelet transform. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE (pp. 6409-6412). IEEE.

[17] Hosseini-Nejad, H., Jannesari, A., Sodagar, A. M., & Rodrigues, J. N. (2015). A 128-channel discrete cosine transform-based neural signal processor for implantable neural recording microsystems. International Journal of Circuit Theory and Applications, 43(4), 489-501.

[18] Hosseini-Nejad, H., Jannesari, A., & Sodagar, A. M. (2014). Data Compression in Brain-Machine/Computer Interfaces Based on the Walsh–Hadamard Transform. IEEE transactions on biomedical circuits and systems, 8(1), 129-137.

[19] Li, N., & Sawan, M. (2015, April). High compression rate and efficient spikes detection system using compressed sensing technique for neural signal processing. In Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on (pp. 597-600). IEEE.

[20] Zhang, J., Suo, Y., Mitra, S., Chin, S. P., Hsiao, S., Yazicioglu, R. F., ... & Etienne-Cummings, R. (2014). An efficient and compact compressed sensing microsystem for

implantable neural recordings. IEEE transactions on biomedical circuits and systems, 8(4), 485-496.